

# Data Clustering for Anomaly Detection

Jose F. Nieves

Polytechnic University of Puerto Rico  
Research Alliance in Math and Science

Computer Science and Engineering Division, Oak Ridge National Laboratory

Mentor: Yu (Cathy) Jiao, Ph.D.

[http://info.ornl.gov/sites/rams09/j\\_nieves\\_rodriguez/Pages/default.aspx](http://info.ornl.gov/sites/rams09/j_nieves_rodriguez/Pages/default.aspx)

## Network intrusion detection

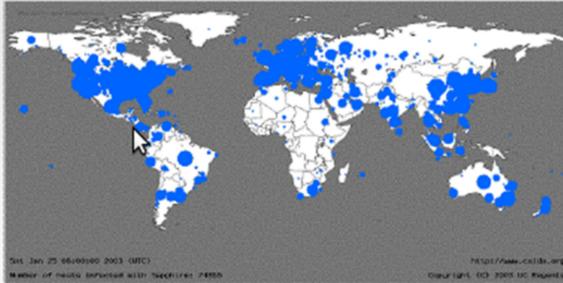


Figure 1. Worm spread after 30 min. [www.caida.org](http://www.caida.org)

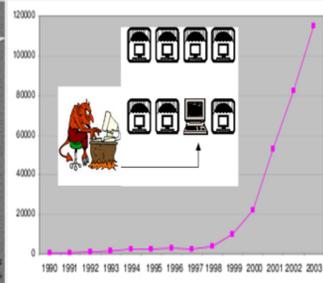


Figure 2. Computer incidents report.

- Unsupervised methods
- Supervised methods

## Clustering-based intrusion detection

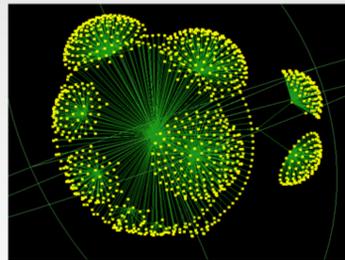


Figure 2. Hierarchical clustering.

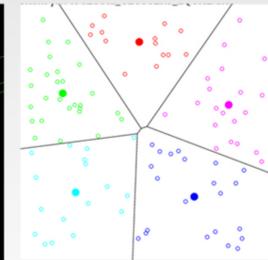


Figure 3. Partitional clustering.

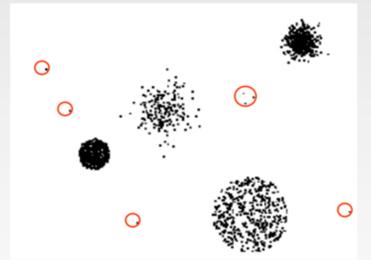


Figure 4. Anomalies in red circles.

- Group unlabeled samples into clusters based on similarities
- Attacks are deviation from 'normal' clusters
- Detect known and unknown attacks

## Software

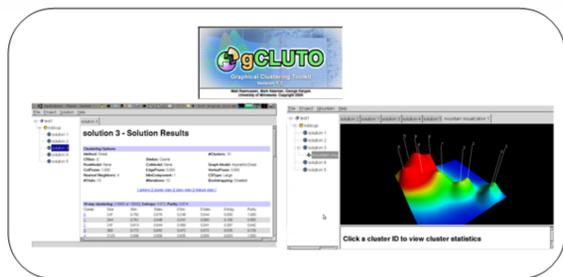


Figure 5. Cluto software for data clustering.

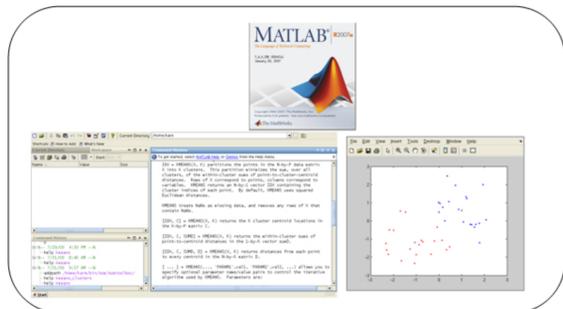


Figure 6. Matlab programming language.

## Data Clustering Steps

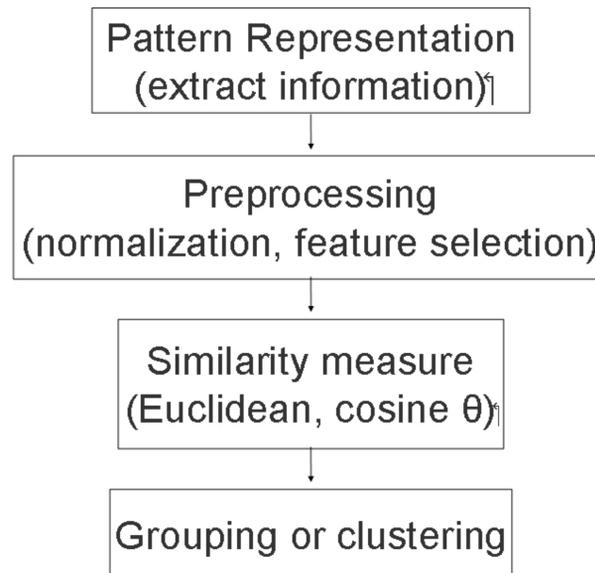


Figure 7. Basic steps for data clustering

## Algorithms

### K-means

- Fast and easy to implement
- Partition data into clusters

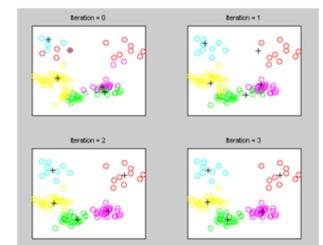


Figure 8. K-means clustering.

- Others
- Fuzzy C-means, SOMs

## Test Data

- Kddcup 1999
- Samples = 9,200
- Features = 80
- 'Attacks' patterns = 2%
- DoS and Probe attacks

## Results

### Bar plot

- Shows the detection rate, false alarm rate, and time for different number of clusters
- K= number of clusters

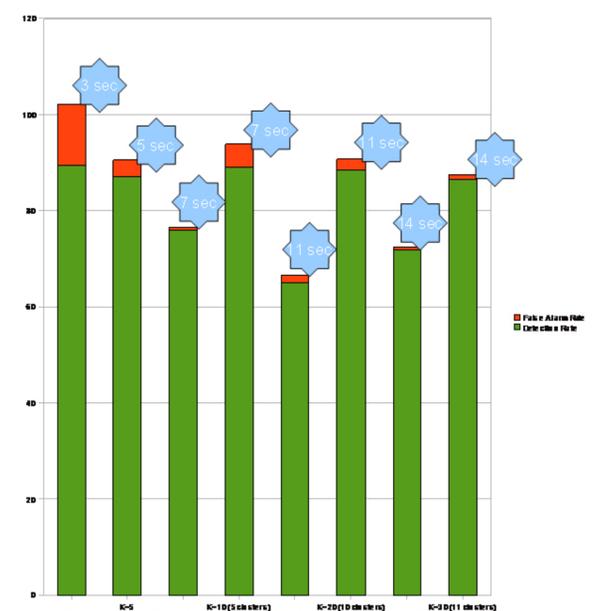


Figure 9. Detection rate and False alarm rate plot.

## References

- Varun Chandola, Arindam Banerjee, and Vipin Kumar. *Anomaly Detection: A Survey*. 2007.
- Leonid Portnoy, Eleazar Eskin, Sal Stolfo. *Intrusion Detection with Unlabeled Data using Clustering*. 2001.
- S Zhong, TM Khoshgoftar, N Seliya. *Clustering-based Network Intrusion Detection*. 2007.